**Identificación y procesamiento de datos atípicos - Outliers**

En esta actividad se obtienen los estadísticos característicos de cada parámetro hidroclimatológico en cada estación, y se identifican, excluyen y completan los valores atípicos, a través de métodos estadísticos.

Para el desarrollo de esta actividad se utilizarán los siguientes métodos:

Método 1, Rango intercuartílico o I Q R: aplica para datos que no siguen una distribución normal y permite identificar valores atípicos a partir de valores que se encuentran por fuera del rango intercuartílico, Q1 a Q3. El valor del primer cuartil puede corresponder al percentil 25, y el valor del tercer cuartil puede corresponder al percentil 75 del conjunto de datos, por lo que el rango intercuartílico se calcula como, Q3 menos Q1.

Método 2, Regla empírica o E R: este método, también denominado regla 3-sigma, o regla 68, 5, 99.7, correspondientes a una distribución normal, permite identificar valores atípicos cuyas observaciones se encuentran fuera de la banda, miu mas o menos k por sigma.

Método 3, núcleo estándar o Z score: este método, similar o idéntico al método 2, permite calificar cada observación a partir del valor, Z = x menos miu sobre sigma. Valores fuera de un valor establecido, p.ej. Z-score=3, son identificados como atípicos.

*Tenga en cuenta que, en este video se realiza una explicación general del procedimiento a realizar, y que el procedimiento detallado para completar las actividades requeridas, se encuentra en la guía de clase disponible en el enlace de la descripción.*

**Objetivos**

* Obtener estadísticas generales de cada estación para cada parámetro hidroclimatológico.
* Identificar valores atípicos a través de diferentes métodos estadísticos.
* Representar gráficamente las series de cada parámetro visualizando también los datos atípicos identificados.
* Excluir valores atípicos de la matriz de estaciones para cada parámetro.
* Reemplazar valores atípicos por valores límite definidos a partir de un rango de confianza.
* Imputar valores atípicos con el valor de la media de cada estación.

**Procedimiento general**

El diagrama mostrado en pantalla, contiene el procedimiento general para la identificación de valores atípicos a partir del método 1, y los procesos derivados de exclusión, ajuste o reemplazo. Para los demás métodos, el procedimiento es similar, y solo cambia el proceso de identificación de valores atípicos.

Para iniciar, descargue el script Outlier punto py, disponible en la carpeta punto .src de este repositorio, y guárdelo en la carpeta local punto src de su equipo.

**Funcionalidades del script**

* Identificación de atípicos por 3 métodos estadísticos.
* Permite definir la tabla dinámica del parámetro hidro climatológico a evaluar.
* El usuario puede excluir estaciones del análisis a través de la variable, station exclude.
* Definición del cuartil inferior q 1 \_ val, y superior q 3 \_ val, que define el rango de exclusión en el método de rango intercuartílico, I Q R.
* Definición manual del multiplicador cap multiplier, o K sigma, que permite definir los valores de reemplazo
* Definición manual del límite de exclusión, z score threshold, en el método de exclusión por núcleo estándar.
* Análisis masivo de estaciones por parámetro hidro climatológico con estadísticos, parámetros de evaluación y gráficas con marcado de atípicos.
* Generación de reportes detallados Mark down, por cada parámetro hidro climatológico evaluado.
* Para cada método y cada parámetro hidro climatológico analizado, crea las siguientes tablas: datos atípicos identificados, datos de entrada sin datos atípicos - drop, datos de entrada con datos atípicos reemplazados – cap, y datos de entrada con datos atípicos imputados - impute.

En el Método 3 o núcleo estándar, Z score, se genera una tabla adicional para cada parámetro hidro climatológico, que contiene los puntajes, a partir de los cuales se realiza la identificación de valores atípicos.

En pantalla, se muestran las instrucciones iniciales del script, en la guía de clase encontrará su contenido completo.

En la carpeta local punto datasets de su repositorio local, cree una carpeta con el nombre IDEAM Outlier, y verifique que en la carpeta IDEAM \_ EDA, del mismo directorio datasets, se encuentren las tablas dinámicas de cada parámetro hidro-climatológico, las cuales fueron obtenidas en la actividad anterior.

Desde el editor de texto, abra el script outlier punto py, y defina las variables, pivot table name, que corresponde a la tabla dinámica a procesar, q1 \_ val, correspondiente al cuartil inferior, q3 \_ val, correspondiente al cuartil superior, cap \_ multiplier, correspondiente al multiplicador K sigma, y z score \_ threshold, correspondiente al límite de exclusión en Z score del método 3.

Desde la consola de comandos de Windows, ingrese a la carpeta IDEAM\_Outlier. Luego, ejecute la instrucción mostrada en pantalla, que realizará todo el proceso de análisis.

Durante la ejecución, podrá observar que en la consola se presenta el detalle de los procesos ejecutados y la previsualización de las diferentes tablas y gráficos.

Luego de la ejecución, podrá observar que en la carpeta local IDEAM \_ Outlier, se han generado diferentes archivos de resultados por cada método, para la tabla de datos de precipitación mensual total.

Durante el proceso de ejecución del script, se genera automáticamente un reporte científico integrado de resultados en formato Mark down, que contiene los resultados mostrados en pantalla.

Al inicio del reporte Mark down, encontrará una descripción general del análisis y las librerías utilizadas, también información relacionada con los parámetros definidos.

A continuación, se presenta un resumen de la cabecera y cola del data frame.

Luego, los tipos de datos, el conteo de valores nulos y los estadísticos característicos iniciales de cada estación.

Para cada método, podrá encontrar una descripción general y los parámetros evaluados para cada estación.

Para cada método, se presenta una gráfica con todos los registros de la serie de datos y los valores atípicos identificados.

También podrá consultar las tablas de valores generadas, relacionadas con los valores atípicos identificados, la tabla dinámica excluyendo los valores atípicos, la tabla con valores reemplazados, y la tabla con valores imputados.

Luego, se presentan los resultados de los nuevos estadísticos de las series con valores reemplazados o imputados.

Para conocer los resultados de los análisis obtenidos a través del método 2, correspondiente a la regla empírica, y los resultados del método 3 o z – score, consulte la guía de clase de esta actividad y los reportes Mark down generados.

En la tabla mostrada en pantalla, se presenta un resumen de los análisis realizados, que incluye el número de estaciones, el número de registros procesados para cada parámetro, y el conteo de valores atípicos identificados.

Al revisar los estadísticos característicos, por ejemplo, de la estación 15 0 1 50 20, correspondiente a datos de precipitación total mensual, podrá observar valores similares de media y desviación estándar:

La conclusión general, es que de acuerdo con los valores atípicos identificados para cada variable hidro-climatológica y evaluando las gráficas compuestas donde se representan todas las series de las estaciones objeto de estudio, se puede evidenciar y concluir que no existen en los conjuntos de datos, valores que deban ser necesariamente excluidos, reemplazados o imputados por métodos estadísticos. Para el desarrollo de las actividades posteriores, podrá trabajar con los datos originales o con las tablas de datos con valores atípicos limpiados y o ajustados, toda vez que se mantienen similares los estadísticos característicos.

**Actividades complementarias**

*En la guía de clase, se encuentran listadas las actividades adicionales que los estudiantes deben desarrollar y documentar para complementar los conocimientos y alcances definidos en este curso. También encontrarás información y referencias adicionales, te recomiendo realizar estas lecturas, que te ayudarán a fortalecer y complementar tus conocimientos.*

*Para completar la Identificación y procesamiento de datos atípicos - Outliers, consulta la guía de clase detallada de esta actividad. Si necesitas ayuda, da clic en el enlace Ayuda o Colabora, que se encuentra en el enlace adjunto de la descripción.*